SYSTEMS AND METHODS FOR LABELING CLUSTERS OF DOCUMENTS

RELATED APPLICATION

[0001] This application is related to U.S. Application Serial No. 10/_____ (Docket No. 02-4034), entitled "SYSTEMS AND METHODS FOR INTERACTIVE CLUSTERING OF DOCUMENTS," filed concurrently herewith and incorporated herein by reference.

[0002] This application claims priority under 35 U.S.C. § 119 based on U.S. Provisional Application No. 60/419,214, filed October 17, 2002, the contents of which are incorporated herein by reference.

GOVERNMENT CONTRACT

[0003] The U.S. Government may have a paid-up license in this invention and the right in limited circumstances to require the patent owner to license others on reasonable terms as provided for by the terms of Contract No. N66001-00-C-8008 awarded by the Defense Advanced Research Projects Agency.

BACKGROUND OF THE INVENTION

Field of the Invention

[0004] The present invention relates generally to multimedia environments and, more particularly, to systems and methods for labeling clusters of similar documents.

Description of Related Art

[0005] When trying to organize large collections of documents, it is sometimes useful to organize these documents into similar groupings, where similarity is determined by some metric,

such as the topics of the documents or their relevance to a particular event. Conventional systems typically receive streams of documents and group the documents into clusters that ideally concern a single event, or more typically, a single topic.

[0006] One particular conventional system includes an event or topic detection system that uses natural language techniques to make a decision about each of the documents it receives. The decision involves the determination of whether a particular document relates to a new event (or topic) that the system has not seen before or an existing event (or topic) that the system has seen before. If the document relates to a new event, then the system creates a new cluster and assigns the document to this new cluster. If the document, instead, relates to an existing event, then the system assigns the document to an existing cluster relating to the event.

[0007] The system usually operates based on a set of rules. One rule is that a document can only be assigned to one cluster. Another rule is that the clusters can only grow and may never be broken. To this effect, the system may never revisit documents that have already been assigned to clusters to determine whether the documents should have been assigned to different clusters.

[0008] The conventional system usually presents the clusters to an end user with no labeling other than, possibly, the number of documents in the clusters. This is of limited usefulness to a user looking for a document in one of the clusters.

[0009] As a result, there is a need for a labeling scheme that creates cluster labels that are indicative of the documents in the clusters and are meaningful to an end user.

SUMMARY OF THE INVENTION

[0010] Systems and methods consistent with the present invention address this and other needs by creating labels for clusters based on document topics that are associated with at least half of the documents in the clusters. The topics may be ranked based on the number of documents relating to the corresponding topics. The topics may then be presented in rank order as labels for the clusters.

[0011] In one aspect consistent with the principles of the invention, a system that generates labels for clusters of documents is provided. The system identifies topics associated with the documents in the clusters and determines whether the topics are associated with approximately half or more of the documents in the clusters. The system then generates labels for the clusters using the topics that are associated with approximately half or more of the documents in the clusters.

[0012] In another aspect consistent with the present invention, a method of creating labels for clusters of documents is provided. The method includes identifying topics associated with the documents in the clusters; determining whether the topics are associated with at least half of the documents in the clusters; adding ones of the topics that are associated with at least half of the documents in the clusters to cluster lists; and forming labels for the clusters from the cluster lists.

[0013] In yet another aspect consistent with the present invention, a system for creating a label for a cluster of documents is provided. The system is configured to identify topics associated with the documents in the cluster and determine whether the topics are associated with approximately half or more of the documents in the cluster. The system is further configured to

rank the topics that that are associated with approximately half or more of the documents in the cluster and generate a label for the cluster using the ranked topics.

[0014] In a further aspect consistent with the present invention, a topic detection system is provided. The topic detection system includes a decision engine and a label engine. The decision engine is configured to receive documents and group the documents into clusters. The label engine is configured to identify topics associated with the documents in the clusters, determine whether the topics are associated with at least half of the documents in the clusters, and form labels for the clusters using the topics that are associated with at least half of the documents in the clusters.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate the invention and, together with the description, explain the invention. In the drawings,

[0016] Fig. 1 is a diagram of a system in which systems and methods consistent with the present invention may be implemented;

[0017] Fig. 2 is an exemplary diagram of the server system of Fig. 1 according to an implementation consistent with the principles of the invention;

[0018] Fig. 3 is an exemplary diagram of the server of Fig. 2 according to an implementation consistent with the principles of the invention;

[0019] Fig. 4 is an exemplary diagram of a portion of the indexing system of Fig. 2 according to an implementation consistent with the principles of the invention;

[0020] Fig. 5 is an exemplary diagram of the event detection system of Fig. 2 according to an implementation consistent with the present invention;

[0021] Fig. 6 is a flowchart of exemplary processing for grouping documents into clusters according to an implementation consistent with the principles of the invention;

[0022] Fig. 7 is a flowchart of exemplary processing for creating a label for a cluster according to an implementation consistent with the principles of the invention; and

[0023] Figs. 8A and 8B are exemplary diagrams of a graphical user interface that may be presented to a user according to an implementation consistent with the principles of the invention.

DETAILED DESCRIPTION

[0024] The following detailed description of the invention refers to the accompanying drawings. The same reference numbers in different drawings may identify the same or similar elements. Also, the following detailed description does not limit the invention. Instead, the scope of the invention is defined by the appended claims and equivalents.

[0025] Systems and methods consistent with the present invention create cluster labels that are indicative of the documents in the clusters and are meaningful to an end user. The labels may be based on document topics that are associated with at least half of the documents in the clusters. The topics may be ranked based on their occurrence in the documents of the cluster. The topics may then be presented in rank order as a label for the cluster.

[0026] In the discussion that follows, a document corresponds to a body of media that is contiguous in time (from beginning to end or from time A to time B). Documents might include

audio documents (e.g., radio broadcasts), video documents (e.g., television broadcasts), and/or text documents (e.g., word processing documents) in any language.

EXEMPLARY SYSTEM

[0027] Fig. 1 is a diagram of an exemplary system 100 in which systems and methods consistent with the present invention may be implemented. System 100 may include clients 110 connected to server system 120 via a network 130. Network 130 may include any type of network, such as a local area network (LAN), a wide area network (WAN), a public telephone network (e.g., the Public Switched Telephone Network (PSTN)), a virtual private network (VPN), or a combination of networks. Clients 110 and server system 120 may connect to network 130 via wired, wireless, and/or optical connections.

[0028] Generally, clients 110 may interact with server system 120 to obtain documents of interest. A user of one of clients 110 may then cause the documents to be automatically grouped into clusters on demand. A client 110 may include a personal computer, a laptop, a personal digital assistant, or another type of device that is capable of interacting with server system 120 to obtain documents of interest. A client 110 may present the documents to a user via a graphical user interface (GUI), possibly within a web browser window.

[0029] Generally, server system 120 may process and maintain documents. Server system 120 may receive documents in a wide variety of formats (e.g., audio, video, and text) and process the documents to extract features and other relevant information from the documents. Server system 120 may also group documents into clusters and, when requested, provide documents to clients 110.

[0030] Fig. 2 is an exemplary diagram of server system 120 according to an implementation consistent with the principles of the invention. Server system 120 may include a server 210, an indexing system 220, an event detection system 230, and a database 240 connected via a network 250. Network 250 may include a LAN, WAN, the Internet, network 130, or other types of direct or indirect connections.

[0031] Server 210 may include a computer or another type of device capable of interacting with clients 110. In one implementation consistent with the principles of the invention, server 210 includes indexing system 220 and/or event detection system 230.

[0032] Fig. 3 is an exemplary diagram of server 210 according to an implementation consistent with the principles of the invention. Server 210 may include bus 310, processor 320, main memory 330, read only memory (ROM) 340, storage device 350, input device 360, output device 370, and communication interface 380. Bus 310 permits communication among the components of server 210.

[0033] Processor 320 may include any type of conventional processor or microprocessor that interprets and executes instructions. Main memory 330 may include a random access memory (RAM) or another type of dynamic storage device that stores information and instructions for execution by processor 320. ROM 340 may include a conventional ROM device or another type of static storage device that stores static information and instructions for use by processor 320. Storage device 350 may include a magnetic and/or optical recording medium and its corresponding drive.

[0034] Input device 360 may include one or more conventional mechanisms that permit an operator to input information to server 210, such as a keyboard, a mouse, a pen, voice recognition

and/or biometric mechanisms, etc. Output device 370 may include one or more conventional mechanisms that output information to the operator, including a display, a printer, a speaker, etc. Communication interface 380 may include any transceiver-like mechanism that enables server 210 to communicate with other devices and/or systems. For example, communication interface 380 may include mechanisms for communicating with another device or system via a network, such as network 250 or network 130.

[0035] As will be described in detail below, server 210, consistent with the present invention, may interact with clients 110, event detection system 230, and/or database 240 to provide documents of interest. Server 210 may perform these tasks in response to processor 320 executing sequences of instructions contained in, for example, memory 330. Alternatively, hardwired circuitry may be used in place of or in combination with software instructions to implement processes consistent with the present invention. Thus, processes performed by server 210 are not limited to any specific combination of hardware circuitry and software.

[0036] Returning to Fig. 2, indexing system 220 may receive document data, including real time data, in a variety of formats (e.g., audio, video, and text), process the data to extract features and other relevant information from the documents, and record the date and time at which the documents were created. In one implementation consistent with the principles of the invention, indexing system 220 may include mechanisms, such as the ones described in John Makhoul et al., "Speech and Language Technologies for Audio Indexing and Retrieval," Proceedings of the IEEE, Vol. 88, No. 8, August 2000, pp. 1338-1353, which is incorporated herein by reference.

[0037] Fig. 4 is an exemplary diagram of a portion of indexing system 220 according to an implementation consistent with the principles of the invention. The portion of indexing system

220 shown in Fig. 4 operates upon audio documents. Indexing system 220 may include similar or dissimilar mechanisms for operating upon other types of media, such as video and text.

[0038] As shown in Fig. 4, indexing system 220 includes audio classification logic 410, speech recognition logic 420, speaker clustering logic 430, speaker identification logic 440, name spotting logic 450, topic classification logic 460, and story segmentation logic 470. Audio classification logic 410 may distinguish speech from silence, noise, and other audio signals in input audio data. For example, audio classification logic 410 may analyze each thirty second window of the input data to determine whether it contains speech. Audio classification logic 410 may also identify boundaries between speakers in the input stream. Audio classification logic 410 may group speech segments from the same speaker and send the segments to speech recognition logic 420.

[0039] Speech recognition logic 420 may perform continuous speech recognition to recognize the words spoken in the segments that it receives from audio classification logic 410. Speech recognition logic 420 may generate a transcription of the speech using a statistical language model. Speaker clustering logic 430 may identify all of the segments from the same speaker in a single document and group them into speaker clusters. Speaker clustering logic 430 may then assign each of the speaker clusters a unique label. Speaker identification logic 440 may identify the speaker in each speaker cluster by name or gender.

[0040] Name spotting logic 450 may locate the names of people, places, and organizations in the transcription. Name spotting logic 450 may extract the names and store them in a database. Topic classification logic 460 may use a probabilistic Hidden Markov Model (HMM) to assign topics to the transcription. In one implementation consistent with the present invention, topic

classification logic 460 uses a technique similar to the one described in John Makhoul et al., "Speech and Language Technologies for Audio Indexing and Retrieval," Proceedings of the IEEE, Vol. 88, No. 8, August 2000, pp. 1338-1353, which was previously incorporated by reference. Topic classification logic 460 may generate a rank-ordered list of all possible topics and corresponding scores for the transcription.

[0041] Story segmentation logic 470 may change the continuous stream of words in the transcription into document-like units with coherent sets of topic labels and other document features generated or identified by the components of indexing system 220. This information may constitute metadata corresponding to the input audio data. Story segmentation logic 470 may store the metadata in database 240.

[0042] Returning to Fig. 2, event detection system 230 may group documents into clusters based on events or topics to which the documents relate. Fig. 5 is an exemplary diagram of event detection system 230 according to an implementation consistent with the principles of the invention. Event detection system 230 may include a decision engine 510 and a label engine 520. The decision engine 510 may include a conventional event or topic detection system, such as the Topic Detection Tracking system developed by the University of Massachusetts, Amherst, as described in J. Allan et al., "UMass at TDT2000," November 2000, pages 109-115.

[0043] Decision engine 510 may include logic that receives a stream of documents over time from, for example, indexing system 220 and/or server 210, and determines, for each of the documents, whether the document is related to an event or topic that decision engine 510 has seen before. If the document is related to a new event or topic (i.e., one that has not yet been seen by decision engine 510), then decision engine 510 may create a new cluster relating to the

event or topic and assign the document to the new cluster. If the document is, instead, related to an existing event or topic, then decision engine 510 may assign the document to an existing cluster that is also related to the event or topic.

[0044] Decision engine 510 may follow the same rules as conventional systems. In other words, decision engine 510 may assign a document to only one cluster. Decision engine 510 may also get only one chance to make a decision about a document and, thereafter, may not change its decision regarding the cluster to which the document is assigned. Decision engine 510 may store its document assignment decisions in an internal memory or, alternatively, in database 240.

[0045] Label engine 520 may include logic that creates labels for the clusters generated by decision engine 510. In another implementation, the functions of label engine 520 are performed by server 210. For each of the clusters, label engine 520 may examine the topics assigned to the cluster documents by indexing system 220. Label engine 520 may then label the cluster with the topics that appear on at least half of the documents in the cluster. The theory is that if a topic does not appear on at least half of the documents in the cluster, then the topic is not representative of the cluster.

[0046] Label engine 520 may rank the topics assigned to a cluster. For example, a topic that is associated with more of the documents in the cluster may be ranked higher than a topic associated with fewer of the documents in the cluster. This ranked list of topics may form a label for the cluster. The clusters with attached labels may be presented to a user via client 110.

[0047] Returning to Fig. 2, database 240 may include a relational database that stores documents from indexing system 220 and, possibly, cluster information from event detection system 230. The contents of database 240 may be accessible to users via clients 110.

EXEMPLARY PROCESSING

[0048] Fig. 6 is a flowchart of exemplary processing for grouping documents into clusters according to an implementation consistent with the principles of the invention. Processing may begin with decision engine 510 receiving a stream of documents over time (act 610). Decision engine 510 may receive the documents from indexing system 220 and/or server 210.

[0049] Decision engine 510 may operate upon the documents to group the documents into clusters (act 620). For example, decision engine 510 may determine, for each of the documents, whether the document relates to a new event (or topic) that decision engine 510 has not seen before or an existing event (or topic) that decision engine 510 has seen before. If the document relates to a new event (or topic), then decision engine 510 creates a new cluster and assigns the document to this new cluster. If the document, instead, relates to an existing event (or topic), then decision engine 510 assigns the document to an existing cluster relating to the event (or topic).

[0050] Label engine 520 may create labels for the clusters generated by decision engine 510 (act 630). Label engine 520 may create a label or reassess a previous label assignment for a cluster on a periodic basis, when a new document is assigned to the cluster, or when cluster information is requested by a user (via client 110).

[0051] Fig. 7 is a flowchart of exemplary processing for creating a label for a cluster according to an implementation consistent with the principles of the invention. Processing may begin with label engine 520 identifying the topics assigned to the documents in the cluster (act 710). In one implementation, label engine 520 obtains the topic information from indexing system 220. In another implementation, label 520 generates the topic information, possibly using

a technique similar to the one described in John Makhoul et al., "Speech and Language Technologies for Audio Indexing and Retrieval," Proceedings of the IEEE, Vol. 88, No. 8, August 2000, pp. 1338-1353, which was previously incorporated by reference. In yet another implementation, label 520 obtains the topic information in some other manner.

[0052] Label engine 520 may then examine each of the topics in the cluster. For example, label engine 520 may determine whether a topic M (where $M \geq 1$) is associated with at least half of the documents in the cluster (act 720). If so, label engine 520 may add topic M to a cluster list (act 730). If topic M is not associated with at least half of the documents in the cluster, label engine 520 may determine whether all of the topics in the cluster have been considered (act 740). If one or more topics have not yet been considered, then label engine 520 may examine the next topic (M + 1), returning to act 720.

[0053] If all of the topics have been considered, then label engine 520 may rank the topics in the cluster list to form a label for the cluster (act 750). For example, label engine 520 may rank a topic that is associated with the majority of the documents in the cluster higher than all other topics. Label engine 520 may rank the topic associated with the next highest majority of the documents in the cluster higher than all other remaining topics, and so on down to one or more topics that are associated with half of the documents in the cluster. Label engine 520 may use this ranked list of topics to form a label for the cluster.

[0054] Label engine 520, or event detection system 230, may store cluster information in database 240. The cluster information, in this case, may include information regarding the clusters to which the documents are assigned and the labels associated with those clusters.

[0055] Returning to Fig. 6, server 210 may present the cluster information to a user upon request (act 640). For example, server 210 may send the cluster information to client 110 for display via, for example, a graphical user interface, such as a browser interface. The cluster information may be presented to the user as a list of clusters that may be sorted based on the number of documents contained in the clusters. In other words, clusters containing larger numbers of documents may be presented higher on the list than clusters containing fewer numbers of documents. The clusters may include assigned labels to make the cluster list meaningful to the user.

[0056] Figs. 8A and 8B are exemplary diagrams of a graphical user interface that may be presented to a user according to an implementation consistent with the principles of the invention. If the user requests to view the clusters generated by event detection system 230, the user may be presented with a graphical user interface, possibly in the form of a browser interface, such as graphical user interface 800 in Fig. 8A. Graphical user interface 800 may include cluster data 810, barchart view option 820, and timeline view option 830.

[0057] Cluster data 810 may include data that identifies the current document count and the current cluster count. The current document count may specify the total number of documents that have been received and processed by event detection system 230. The current cluster count may specify the total number of clusters in which the documents have been grouped.

[0058] Barchart view option 820 and timeline view option 830 are two manners by which the clusters may be presented to the user. In other implementations consistent with the present invention, there are more or fewer ways of presenting the clusters to the user. Barchart view

option 820 may display the clusters in the form of a barchart. Timeline view option 830 may display the clusters in the form of a timeline.

[0059] Fig. 8B is an exemplary diagram of graphical user interface 800 that may be presented when providing clusters in barchart form according to an implementation consistent with the principles of the invention. Graphical user interface 800 may present the clusters as a series of bars, the length of which relate to the number of documents in the clusters. The bars may be sorted by cluster size, with larger clusters being presented first. Each of the bars may have an associated label that corresponds to the label generated for the cluster by label engine 520.

[0060] The user may select one of the bars to view the documents included in the cluster. The documents may then be presented to the user in chronological order (i.e., sorted based on the date and time at which the document was created), with the more recent documents being presented first. In other implementations, the documents are presented in other ways.

CONCLUSION

[0061] Systems and methods consistent with the present invention create labels for clusters of documents, such that the labels are indicative of the documents in the cluster and are valuable to an end user seeking a document in one of the clusters. The labels may be based on document topics that are associated with at least half of the documents in the clusters. The topics may be ranked based on the number of documents with which the topics are associated. The topics may then be presented in rank order as a label for the cluster.

[0062] The foregoing description of preferred embodiments of the present invention provides illustration and description, but is not intended to be exhaustive or to limit the invention to the

precise form disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention.

[0063] For example, it has been described that only topics that are associated with at least half of the documents in the cluster are used for the cluster label. In other implementations, the criteria is changed to include topics associated with more or fewer than half of the documents.

[0064] While series of acts have been described with regard to Figs. 6 and 7, the order of the acts may be varied in other implementations consistent with the principles of the invention. Also, non-dependent acts may be performed in parallel.

[0065] Further, certain portions of the invention have been described as "logic" that performs one or more functions. This logic may include hardware, such as an application specific integrated circuit or a field programmable gate array, software, or a combination of hardware and software.

[0066] No element, act, or instruction used in the description of the present application should be construed as critical or essential to the invention unless explicitly described as such. Also, as used herein, the article "a" is intended to include one or more items. Where only one item is intended, the term "one" or similar language is used. The scope of the invention is defined by the claims and their equivalents.